CO-CHANNEL SPEECH AND SPEAKER IDENTIFICATION STUDY

Robert E. Yantorno
Associate Professor
Electrical & Computer Engineering Department

College of Engineering
Temple University
12th & Norris Streets
Philadelphia, Pa 19122-6077

Final Report for:
Summer Research Faculty Program
Rome Labs

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Speech Processing Lab
Rome Labs
Rome, New York

September 1998

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
00091998	N/A	-

Title and Subtitle CO-Channel Speech and Speaker Identification Study	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Summer Research Faculty Program Rome Labs Rome, NY	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es)	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
The original document contains color images.

Abstract
This study was comprised of two parts. The first was to determine the effectiveness of speaker identification under two different speaker identification degradation conditions, additive noise and speaker interference, using the LPC cepstral coefficient approach. The second part was to develop a method for determination of co-channel speech, i.e., speaker count, and to develop an effective method of either speech extraction or speech suppression to enhance the operation of speaker identification under co-channel conditions. The results of the first part of study indicate that under conditions of the same amount of either noise or corrupting speech, for example 0 dB SNR or TIR (target-to-interference ratio), noise is much more detrimental than corrupting speech to the operation of the speaker identification. For example, with 100% of 0 dB corrupting speech there still occurs a certain number of correct speaker identifications, i.e., about 40% accuracy. Ten (10) dB TIR interfering speech, as well as small amounts of interfering speech, i. e., 40% 0 dB TIR are not as detrimental to speaker identification. The results of the second part of the study indicate that a system for speaker count and speaker separation is possible. The harmonic sampling approach, developed during the study, uses the periodic structure of the fine structure of the frequency characteristics of voiced speech. Successful reconstruction of a single speaker indicates the potential of this approach as a candidate for speech separation. Also, it was shown that detection of co-channel speech is possible using the harmonic sampling approach. Further improvements as well as other possible approaches to the co-channel speech problem are discussed.

Subject Terms

Report Classification	**Classification of this page**
unclassified	unclassified
Classification of Abstract	**Limitation of Abstract**
unclassified	UU

Number of Pages
20

CO-CHANNEL SPEECH AND SPEAKER IDENTIFICATION STUDY

Robert E. Yantorno
Associate Professor
Electrical & Computer Engineering Department
Temple University

Abstract

This study was comprised of two parts. The first was to determine the effectiveness of speaker identification under two different speaker identification degradation conditions, additive noise and speaker interference, using the LPC cepstral coefficient approach. The second part was to develop a method for determination of co-channel speech, i.e., speaker count, and to develop an effective method of either speech extraction or speech suppression to enhance the operation of speaker identification under co-channel conditions. The results of the first part of study indicate that under conditions of the same amount of either noise or corrupting speech, for example 0 dB SNR or TIR (target-to-interference ratio), noise is much more detrimental than corrupting speech to the operation of the speaker identification. For example, with 100% of 0 dB corrupting speech there still occurs a certain number of correct speaker identifications, i.e., about 40% accuracy. Ten (10) dB TIR interfering speech, as well as small amounts of interfering speech, i. e., 40% 0 dB TIR are not as detrimental to speaker identification. The results of the second part of the study indicate that a system for speaker count and speaker separation is possible. The harmonic sampling approach, developed during the study, uses the periodic structure of the fine structure of the frequency characteristics of voiced speech. Successful reconstruction of a single speaker indicates the potential of this approach as a candidate for speech separation. Also, it was shown that detection of co-channel speech is possible using the harmonic sampling approach. Further improvements as well as other possible approaches to the co-channel speech problem are discussed.

CO-CHANNEL SPEECH AND SPEAKER IDENTIFICATION STUDY

Robert E. Yantorno

Introduction

Co-channel speech is defined as a speech signal which is a combination of speech from two talkers. The goal of co-channel research has been to be able to extract the speech of one of the talkers from the co-channel speech. This can be achieved by either enhancing the target speech or suppressing the interfering speech. This co-channel situation has presented a challenge to speech researchers for the past 30 years. There are systems where the separation of two speakers is possible and this is well documented in the literature. However, this requires that there be more than one sensor (in the case of speech, more than one microphone) and therefore, by making use of the dissimilar recording conditions, the speech from two different speakers can be extracted, for example Chang *et al* (1998) and Weinstein *et al* (1993). Some recent investigations conducted on co-channel speaker separation are; Savic et al (1994), Morgan *et al* (1995), Benincasa & Savic (1997), Yen & Zhao (1997) for speech recognition and Meyer *et al* (1997) using the approach of amplitude modulation mapping and reassignment (a vector quantization process). However, separation of the target speaker from co-channel speech has been very difficult. Therefore, to make the problem more manageable, it is worthwhile to ask what is the final use of the target speech. For example, if the final goal is that a human listener will use the speech, then intelligibility and quality would be important characteristics of the extracted speech. However, if the extracted speech is to be used for speaker identification, then one would be concerned with how much and what type of target speech is need to perform "good" speaker identification, i.e., voiced and unvoiced speech or just voiced speech. Therefore, determining the effect of speaker interference on speaker identification would be of considerable interest. Also, the development of an effective target speaker extraction technique, which would provide for major improvement of co-channel speech, would also be a very useful tool. The goal of this study is to better understand how interfering speech degrades the functioning of speaker identification, and to also develop a method for extracting enough target speech from co-channel speech to provide sufficient speech information about the target speaker that one can make a reliable identification of the target speaker.

The situation with co-channel speech can be viewed in three different ways, i.e., as either an extraction of the target speech, as a suppression of the interference speech, or as an estimation of both speech signals. All these methods have been developed and each requires a very different approach. Therefore, study of the effect of speaker interference on speaker identification would be helpful in choosing between extraction, suppression or estimation, and would also provide information on the development of the method to assist in speaker identification under varying co-channel conditions.

Speaker Identification Study

As outlined above, the first part of the study was to determine the effectiveness of speech identification using the LPC cepstral coefficient approach under two different speaker identification degradation conditions, additive noise and speaker interference. Information on that part of the study is discussed below.

Additive Noise

Methodology

For the initial part of the study, speech material was taken from the Timit database. The number of speakers used for training was 15 males and 15 females. The number of files for training for each speaker was 5. The files were taken from the dialect region 1 (dr1 subdirectory). Also, 15 male and 15 female speakers were used for testing, and were taken from the same dr1 subdirectories. It should be noted that the test speakers were the same speakers used for training, i.e., the speaker identification tests were conducted under closed conditions. The files used for training were the “sx” prefix speech files and the testing files were “sa” and “si” prefix speech files, which were all different speakers speaking the same utterance.

Results

The initial part of section 1 of this study was to determine the effect of noise on the accuracy of the speaker identification. The amount of noise added was varied, either by adding a specific amount of noise, in dB, to the entire utterance or by adding a percentage of 0 dB noise to the speech. The range of added noise in dB was 0 to 30 dB, and the percentage range of 0 dB added noise was 20 to 100 percent. Noise was always added to the center of the utterance. This was

done to ensure that even at a low percentage of added noise, i.e., 20% noise, the probability of noise being placed in a region of speech would be greater than if the noise was placed at the beginning or end of the speech file. This was done because for most utterances there is a certain amount of silence prior to the onset of speech. Therefore, if noise was added to the beginning of the speech file, a certain part of the noise would always be added to silence, and would not contribute to the degradation of the speaker identification process. Results of the percent noise added experiments are shown in Figures 1a. and 1b. One important observation that can be made is that there is an almost linear inverse relationship between percent correct speaker identification and percent noise (figure 1b.).

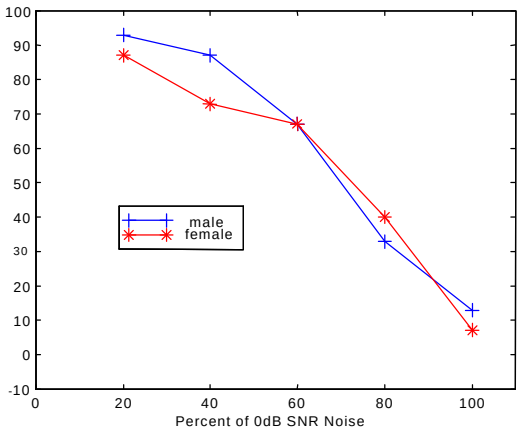

Figure 1a.

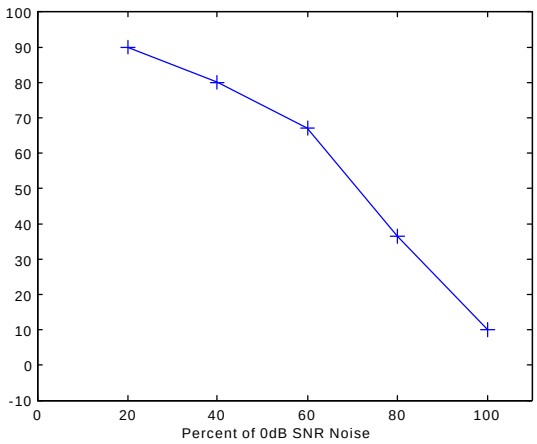

Figure 1b.

Figure 1. Speaker Identification – Percent Correct versus 0 dB SNR of Noise added to speech. Figure 1a. – male and female speakers. Figure 1b. represents the combined results of figure 1a.

The results of varying the signal-to-noise ratio are shown in figures 2a and 2b. The most dramatic decrease in speaker identification occurs between 20 and 10 dB. Also, it can be noted in figure 2b that 10 dB SNR is enough to totally degrade the speaker identification operation. Openshaw & Mason (1994) obtained similar results, where the effects of noise on the speaker identification using both the mel-cepstral coefficient technique as well as perceptual linear prediction-RASTA method were studied.

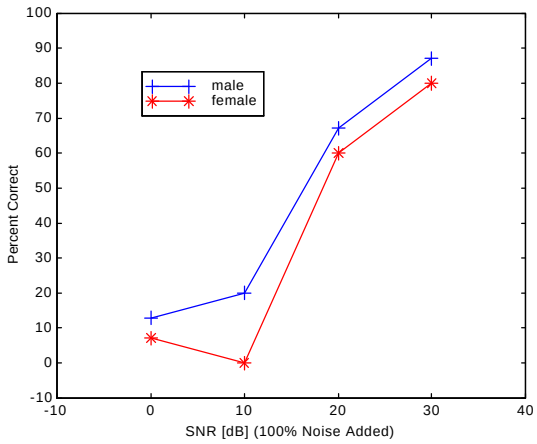

Figure 2a.

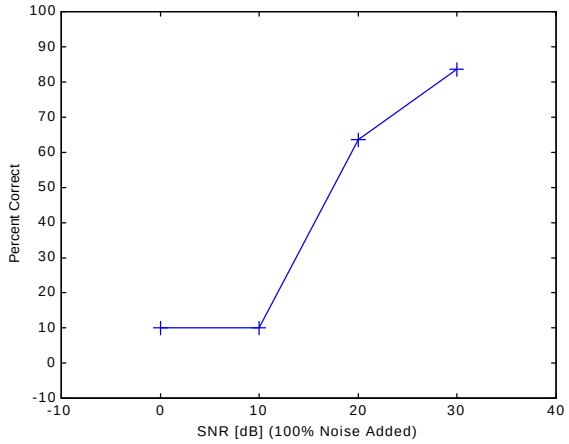

Figure 2b.

Figure 2. Speaker Identification – Percent Correct versus SNR of noise in dB added to speech (100% of speech corrupted by noise). Figure 2a. – male and female speakers. Figure 2b. is the combined result of figure 2a.

If one assumes that the noise experiments represent extreme conditions for speech corrupting speech, then certain conclusions can be drawn about the level, in dB, and amount of 0 dB speech that might be tolerable before one will observe a decease in percent correct speaker identification of co-channel speech. For example, one could tolerate about 40% of 0 dB SNR corruption speech with only a slight decrease of about 15 percent in percent correct (figure 1b.). Also, any corrupting speech of 20 dB target-to-interfere ratio (TIR), also referred to as signal-to-interference ratio (SIR), should have little effect on the speaker identification, in this case about 20 percent decrease in speaker identification percent correct. Therefore, the noise experiments provide a lower bound measure from which one can infer how well the speaker identification system will work under co-channel conditions.

Speaker Interference

Methodology

To determine the effect of corrupting speech on the accuracy of speaker identification, corrupting speech was added to test files. The amount of corrupting speech added was varied either by adding a specific amount of corrupting speech, in dB, to the entire utterance or by adding a percentage of 0 dB target-to-interfere ratio (TIR). The range of corrupting speech in dB was 0 to 30 dB TIR, and the percentage of 0 dB TIR ranged from 20 to 100 percent. As with the noise experiments, and for the same reason, corrupting speech was added to the center of the utterance.

Results

Two sets of four experiments were conducted. For one set, the corrupting speech was drawn from one of the speakers of the training and testing data, but was not the same utterance as used for that speaker's training or testing. These experiments are identified as "closed set" experiments and the results are shown in figures 3a. and 3b.

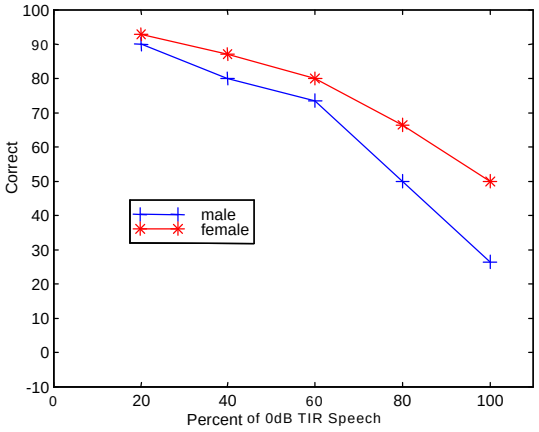

Figure 3a.

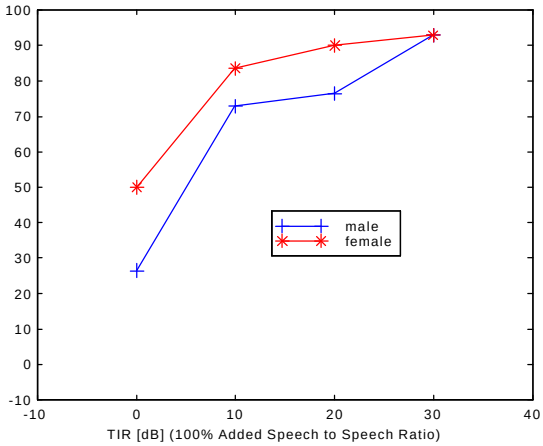

Figure 3b.

Figure 3. "Closed Set" Speaker Identification Experiments. Figure 3a. Percent Correct versus Percent of 0 dB TIR (Target-to-Interferer Ratio). Figure 3b. – Percent Correct versus TIR in dB of corrupting speech added to speech (100% of speech corrupted by speech).

For the other set of experiments, the corrupting speech was drawn from speakers outside the training and testing data. These experiments are identified as "open set", and the results are shown in figures 4a. and 4b. For both the closed and open set experiments there were four different types of experiments: 1.) male speech corrupted by either male or 2.) separately by female speech (results are identified as male in figures 3 and 4) and 3.) female speech corrupted by either male or 4.) separately by female speech (results are identified as female in figures 3 and 4).

One major observation that can be made with respect to figures 3 and 4 is that even with 100% corruption of 0 dB TIR there still exists a certain number of correct speaker identifications, i.e., about 40% accuracy. This indicates that corrupting speech has a smaller effect on the speaker

identification system than does noise, substantiating the point made earlier about noise contamination of speaker identification being the "worst" case.

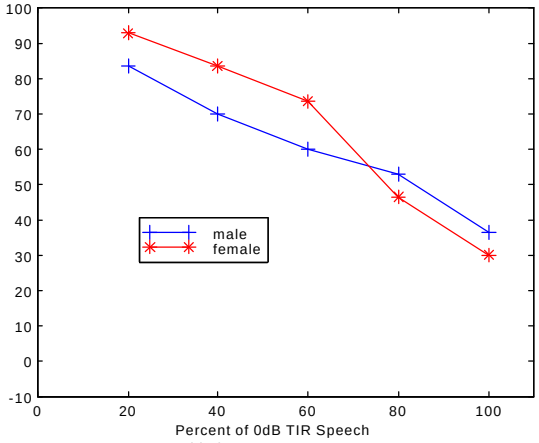

Figure 4a.

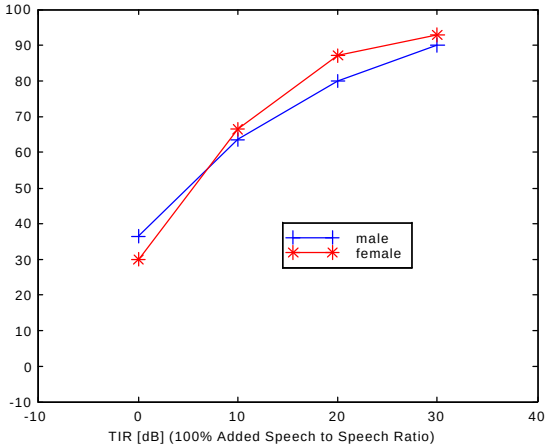

Figure 4b.

Figure 4. "Open Set" Speaker Identification Experiments. Figure 4a. Percent Correct versus Percent of 0 dB TIR. Figure 4b. – Percent Correct versus TIR in dB of corrupting speech added to speech (100% of speech corrupted by speech).

This result seems reasonable because 0 dB TIR does not spread the energy over the entire utterance as in the case of the noise experiments. It is also evident that although there is an almost linear inverse relationship between percent correct and percent added corrupting speech (figure 4a.), as in the case of noise (figure 1b.), the slope is not as steep, again as expected. Also, for the closed set experiments, figures 3a. and 3b., male speaker identification appears to be more sensitive to corrupting speech than does female speaker identification. A smaller effect can be observed with the open set experiments shown in figures 4a. and 4b.

Finally, a comparison is made between the "open" and "closed" set experiments and the results are shown in figures 5a. and 5b. The major observation to be made is that for speaker identification, corrupting speech with speech from outside the training data tends to have a greater effect on the percent correct than corrupting speech from within the training data, i.e., for the 100% of 0 dB TIR corruption speech condition the percent correct was 40% (for the closed condition) and 35% (for the open condition), or about 5% decrease in percent correct.

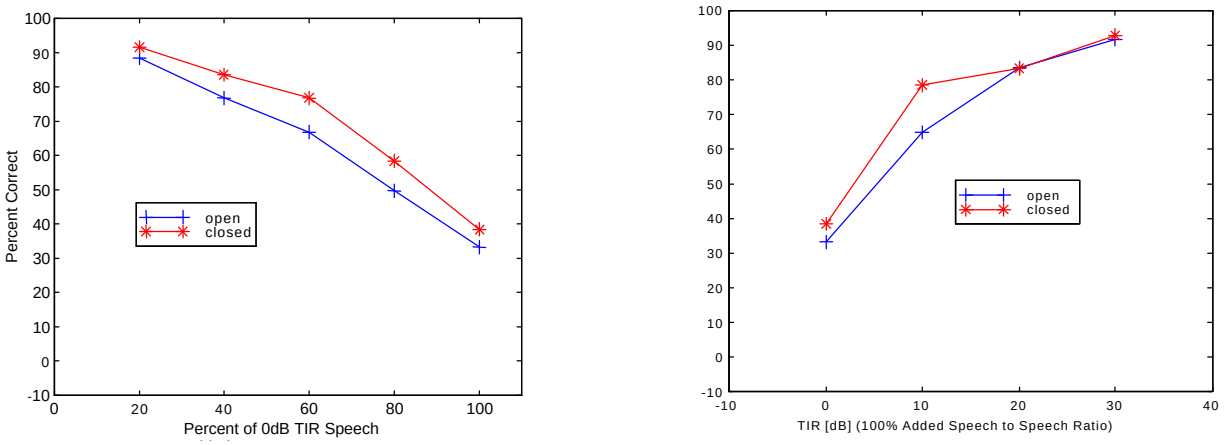

Figure 5. Comparison of data from "Closed Set" and "Open Set" Speaker Identification Experiments. Figure 5a. Percent Correct versus Percent of 0 dB TIR. Figure 5b. – Percent Correct versus TIR in dB of corrupting speech added to speech (100% of speech corrupted by speech).

It should be noted that Yu & Gish (1993) obtained comparable results for experiments similar to the ones shown in figures 3 and 4. Their goal was to identify either one or both speakers engaged in dialog using speech segments rather than frames and speaker clustering.

Co-channel Speech

Introduction

The second part of this project was to develop a method for effective determination of co-channel speech, i.e., speaker count, and be able to extract enough speech information about the target speech so that a reliable identification of that speaker could be made. The method which was developed is based on harmonic sampling in the frequency domain, and is outlined below. It should be noted that the method presented here is similar to the approaches of Doval & Rodet (1991) and Casajus Quiros & Enriquez (1994), which were used for fundamental frequency determination and tracking of music signals. It should also be mentioned that there is some similarity between the method outlined here and the maximum likelihood pitch estimation developed by Wise *et al* (1976). However, Wise *et al* used the autocorrelation and the time domain for their approach, whereas the frequency domain and the magnitude spectrum will be used for the method outlined here. However, they did mathematically analyze their method in the

frequency domain and determined that maximizing their peak estimator was equivalent to finding the comb filter which passes the maximum signal energy, which is the basis for the harmonic sampling method presented in this study.

Harmonic Sampling Method

If one observes the frequency domain characteristics of a voiced portion of a speech signal, it can be noted that there are two distinct attributes, the spectral envelope of the speech signal, and the fine structure which is a series of pulses. The spectral envelope consists of the frequency characteristics of the vocal tract. The fine structure consists of the frequency characteristics of the excitation which is the input to the vocal tract. The excitation for voiced speech is characterized by periodic time pulses produced by the vocal cords which produce periodicity in the frequency domain. The fine structure and its periodicity are illustrated in figure 6 below. The periodicity of the fine structure is the basis for the approach presented. Because we are using the fine structure, and the fine structure only exists during voiced speech, this means that only the voiced portions of speech will be used. Also, voiced speech appears to carry much more speaker identification information than unvoiced speech.

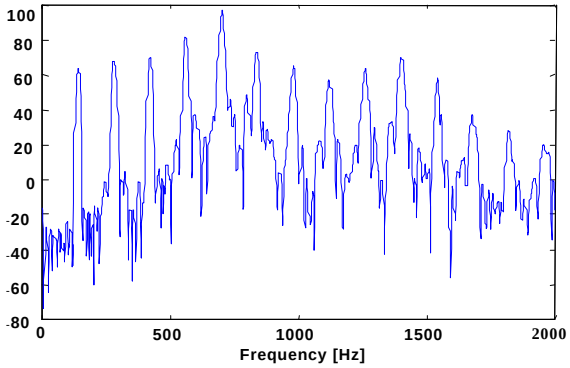

Figure 6. Frequency characteristics of a frame of speech – magnitude in dB versus frequency in Hz, 800 point frame, Hamming windowed, and sampled at 8 kHz.

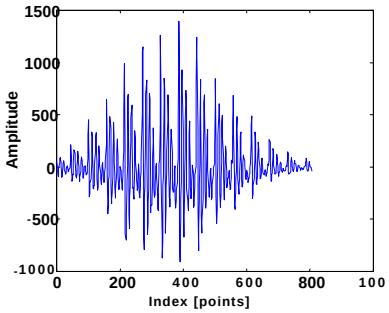

Figure 7a.

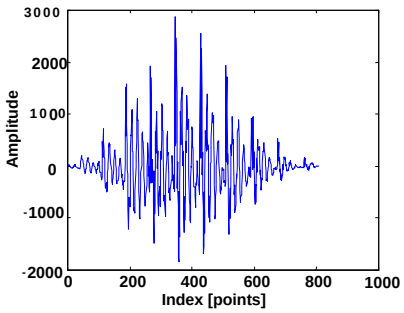

Figure 7b.

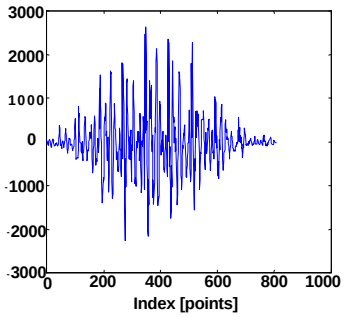

Figure 7c.

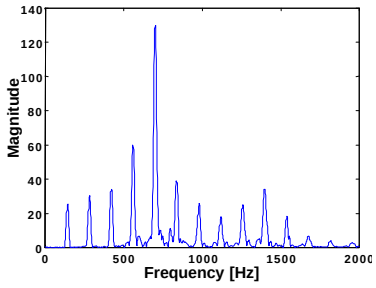

Figure 7d.

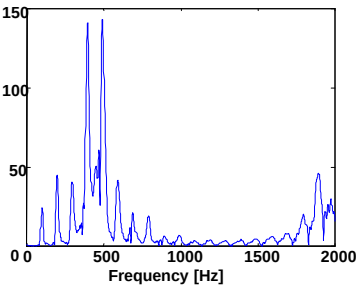

Figure 7e.

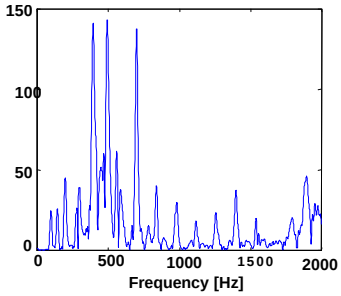

Figure 7f.

Figure 7. Time and frequency characteristics of speech from single speakers - speaker #1 (7a. and d.), speaker #2, (7b. and 7e.) and combined speakers #1 and #2 (7c. and 7f.)

For the case of co-channel speech, it is expected that the frequency characteristics of two speakers is additive. However, the magnitude characteristics are additive but also contain a cross term. Therefore, the overall fine structure will be a quasi-linear combination of fine structure of the two speakers' speech. This is illustrated in figure 7f. As stated, the approach in this study uses the fine structure as a method for determining if there is more than one speaker present and also, very importantly, uses the information obtained as a means of extracting the target speech.

The experiment, for speaker count and speaker separation, entailed using a variable spacing inverse comb filter to sample the magnitude spectrum. If the spacing between the filter lines and between the first line and the vertical axis are variable then one has a tunable comb filter. Therefore, there should be a maximum when the spacing between the comb lines is equal to the fundamental frequency of the speech, as discussed previously and stated by Wise *et al* (1976) for

their pitch detection method. Using the harmonic sampling method, the frequency spectrum is "swept", sampling the spectrum at discrete frequencies, and adding all of samples, in this case 31, at each frequency step. The result will be a graph with a peak at the fundamental frequency. However, after some work on filter design and further considerations, it was recognized that one need only sample the spectrum, and therefore, the comb filter was not needed. A series of diagrams of harmonic sampling for various values of sampling spacing are shown in figure 8 below.

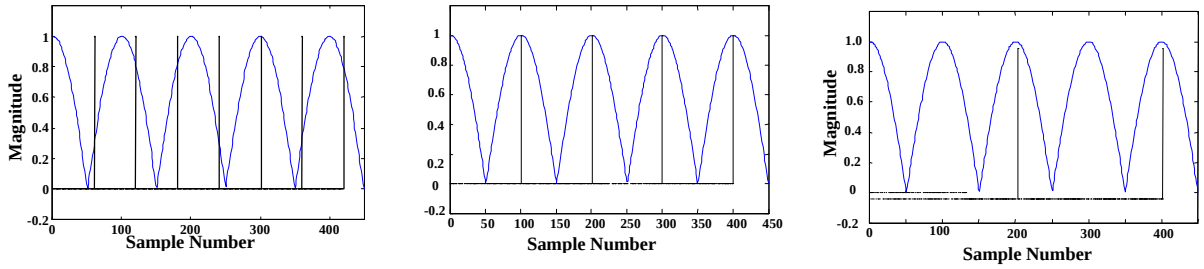

Figure 8a. Figure 8b. Figure 8c.

Figure 8. Harmonic sampling. Three different harmonic sampling spacings, less than fundamental (8a), at fundamental (8b), and at 2x fundamental (8c).

The sampling of the magnitude spectrum results in a peak at the fundamental, but also results in "harmonic-like peaks" at locations related to the fundamental. This situation is illustrated in figure 8c. It is evident that there will be a peak due to the spacing as illustrated in figure 8c. Because we are using a fixed number of lines, the height of that peak will be about half the height of the main peak. However, there will also be a peak at half the fundamental, for example, using figure 8b., if there were twice as many lines as shown, there would be lines located halfway between the lines shown, at the nulls of the spectrum.

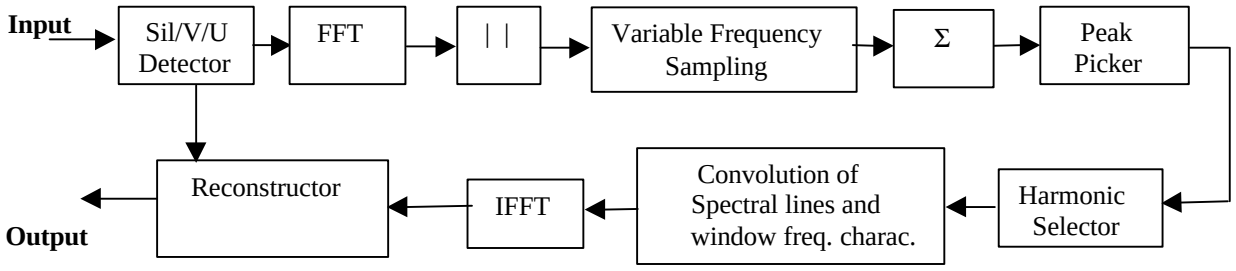

Figure 9. Block diagram of target extraction procedure. Input is co-channel speech and output is target speech.

The height of this peak will also be about one-half the height of the main peak. The entire harmonic sampling method is outlined in the block diagram, shown in figure 9 above.

A peak-picking algorithm was developed to determine the main peak associated with the stronger speaker's speech. To illustrate the effectiveness of the peak-picking algorithm uncorrupted speech was used. The spectrum of uncorrupted speech was sampled at the fundamental and all of the harmonics up to harmonic 30. A plot of the magnitude spectrum, the sampled harmonic spectrum, and the reconstructed magnitude spectrum are shown in figures 10a., 10b. and 10c. respectively.

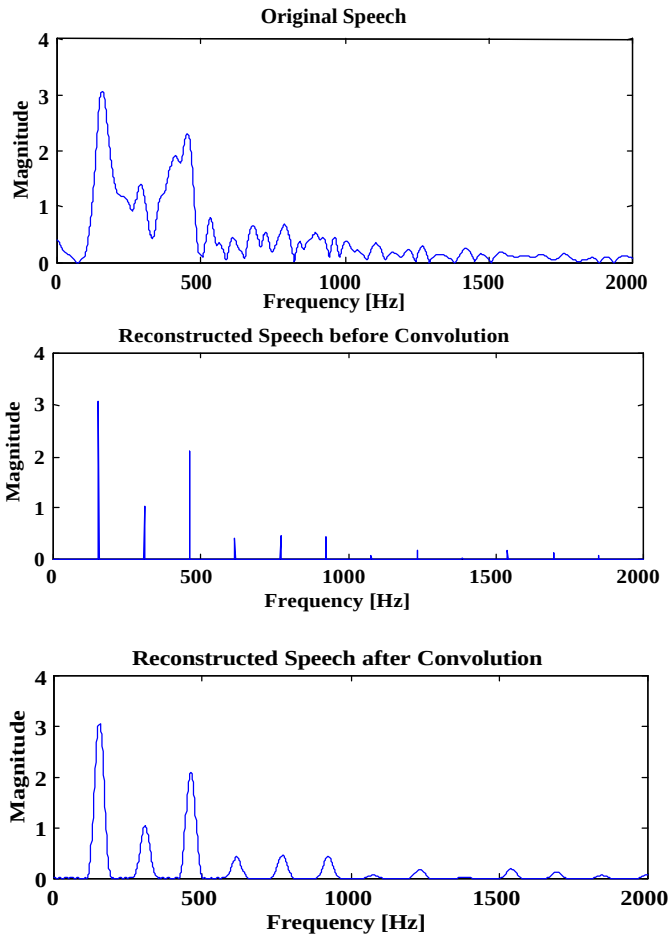

Figure 10. Magnitude frequency plots for original (upper), harmonically sampled (middle) and reconstructed spectrum (lower).

It was determined that the magnitude spectrum provided a much better main peak, in terms of height above the other random peaks, than either the power spectrum or log magnitude spectrum. Also, a slight positive slope was observed for the harmonic sampling results. A fourth order polynomial fit of the data was subtracted from the data to provide a better environment for peak picking. Windows of various lengths were investigated, and a length of 400 points appeared to be the optimal in terms of frequency-time resolution and shape and size of the major harmonic peak. The 400-point frame was windowed using a Hamming window and was zero padded to 8000 points prior to performing the FFT. This results in each point in the harmonic sampling result plot to be equal to 1 Hz. It should be noted that this is not the resolution of the harmonic sampling approach. Note, the Timit speech data was down sampled prior to processing from 16 KHz to 8 KHz using Entropic's ESPS sfconvert utility. Fifty percent overlap was used during the analysis phase to compensate for the windowing of the frame.

Once the spectral lines are obtained, the convolution of these lines with the frequency characteristics for the Hamming window is necessary in order to obtain a "window" time function similar to the original speech frame from which the spectral data was obtained (see figure 10 above). Finally, for reconstruction, the frames were overlapped by 50% to duplicate the overlap process used for extracting and analyzing the speech frame. Using both the harmonically sampled magnitude and phase characteristics for reconstruction did not provide a very good reproduction of the new frame. Therefore, the entire phase characteristics were used for reconstruction. This results in very good duplication as shown in figure 12 below.

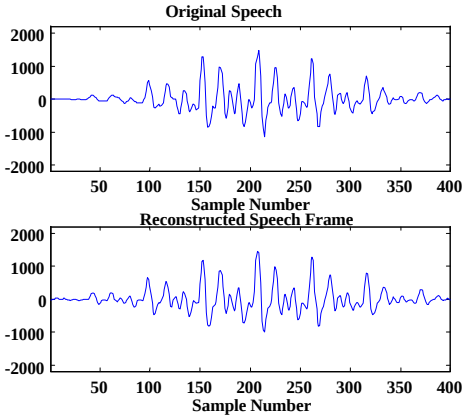

Figure 11. Windowed frame of speech, original speech (upper) and reconstructed speech (lower).

How well does this approach track the pitch of a single speaker? A voiced/unvoiced detector was obtained from Dan Benincasa and Stan Wenndt (of the Speech Processing Lab). The result, for male speech, is shown in the figure 12 below. It is evident that the algorithm works very well, and therefore this is a good candidate for use as a pitch tracker as well as a tool for speech extraction.

Figure 12a. (upper) and 12b. (lower) Figure 12c. (upper) and 12d. (lower)

Figure 12. Time plots of speech signals. Figure 12a. Identification of voiced (1), unvoiced (0.5) and silence (0) sections of speech signal. Figure 12b. Pitch versus time for voiced portions of utterance. Figures 12c. & 12d. Original and reconstructed speech, respectively.

How effective is the algorithm in determining the existence of two speakers? The results shown in figures 13a., 13b. and 13c. represent the results of speech data shown in figures 7a., 7b., and 7c., respectively for single speakers (figures 13a. and 13b.) and co-channel speech (figure 13c.). As can be noted, the pitch of both speakers is clearly seen in figure 13c., as marked by the straight lines in the middle of the two tallest peaks. Note the peak on the right in figure 13c is not at the location where one would expect a multiple of the pitch of the largest peak, and therefore the peak on the right represents the pitch of another speaker.

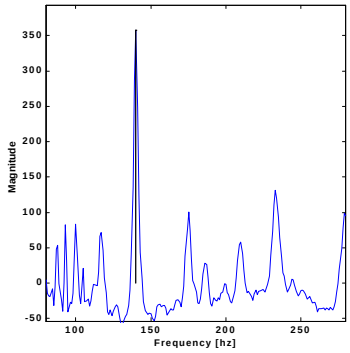

Figure 13a.

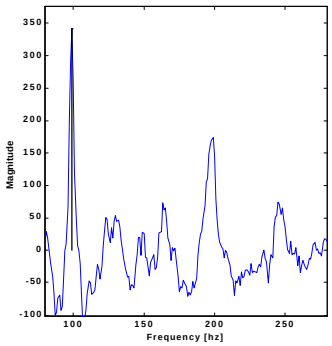

Figure 13b.

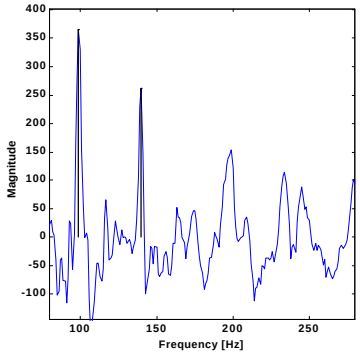

Figure 13c.

(speaker #2) figure 13b., and for co-channel speech (a mixture of speaker #1 and speaker #2) figure 13c.

Conclusions

It has been shown that the harmonic sample approach can be used for speaker count, where the pitch of both speakers are not the same. It has also been shown that the harmonic sampling approach can be used to extract very effectively and quite accurately the voiced portions of a single speaker. The harmonic sampling approach outlined above offers promise as a mechanism for extracting the voiced portion of target speech. However, there are problems to overcome and improvements which can be made.

Further Improvements for the Harmonic Sampling Method

For example, the occurrence of small spurious peaks close to the main pitch peak need to be eliminated, possibly by smoothing the harmonic sampling resultant or by using their occurrence as an indicator not to use that frame of speech. Also, large spurious peaks far from the main peak need to be understood in terms their origin, and to be reduced, eliminated or possibly used as an indicator not to use that frame of speech. Also, the harmonic sampling method may be able to be used to determine if the frame being analyzed is voiced, unvoiced or silence, possibly by using some sort of indicator of the existence of large peaks for voiced, small peaks for unvoiced and no peaks for silence. The final item for investigation is to determine a way to identify speaker #1 speech and speaker #2 speech, and to track their speech. This will require more sophisticated ways in which the speech of a specific speaker is identified. One possibility would be to use a pitch tracking approach such as the one developed by Doval & Rodet (1993) who

used a probabilistic Hidden Markov model or model space, or the approach of Dorken & Nawab (1994) which uses principle decomposition analysis.

It should be noted that the approach outlined here might be useful as a pitch detector as well as a voiced/unvoiced detector. The method of Wise *et al* (1976) has been shown to be resistant to noise. Therefore, because of the similarity between harmonic sampling and maximum likelihood pitch estimation, harmonic sampling should also be resistant to noise.

Finally, this speech separation approach shows promise in terms of being able to extract the interfering speech by subtracting the reconstructed speech from the original speech using either the frequency domain or the time domain.

Areas of Possible Further Study

Speaker Count

I feel that developing a system for identifying co-channel speech is possible. However, to be able to identify co-channel speech will require using unorthodox types of approaches. One such approach would be to use the approach of speech recognition similar to that used for language identification. It seems reasonable that co-channel speech, which is the result of speech corrupted by speech, will not have the same time domain structure as traditional speech, and therefore will not have the same type of phonetic structure as single speaker speech. This means that recognition would not be successful and would therefore indicate the existence of co-channel speech.

It also seems possible that an effective speaker count approach could be developed using information in the time domain. It is evident from inspection of the co-channel speech that there is a dramatic change in the overall structure as compared with single speaker speech (compare figure 7c. with either figure 7b. or 7a.). If a time domain speaker count system could be made then this speaker count system could be used as the front-end of a speech separation system. This would ascertain whether a frame of speech is from a single or multiple speakers. If it is co-channel speech then it would be processed by the speaker separation system. However, if it is

from a single speaker the frame would not be processed, thereby reducing computation time as well as eliminating any possible degradation of the speech using the speech separation system.

Another possible approach would be to use LPC to determine the presence of co-channel speech. For example, if two speakers were talking at the same time, they usually would not be saying the same thing. Therefore, each speaker would be producing different speech sounds and each speaker would then have a different vocal tract configuration at any instant in time. This would seem to suggest that a series of LPC analyses could be done on a single frame. Assuming that one has co-channel speech, LPC analysis would be performed on the speech. Once a set of LPC coefficients had been obtained their effect could be subtracted from the co-channel speech by inverse filtering. Then performing a subsequent LPC analysis on the inverse filtered signal should produce another set of LPC coefficients only if co-channel speech is present. Note, this approach could only be used to detect co-channel speech. It would not be able to extract the target speaker's speech.

Speaker Separation

Although there is no information available about using singular value decomposition (SVD) for co-channel speech, it might be a possible approach. Kanjilal & Palit (1994) have used SVD as a means of extracting two periodic stationary waveforms from noisy data; in this case the waveforms were maternal and fetal electrocardiograms. It should be noted that their approach had no requirement for multiple sensors, but did require that the signals be stationary.

Finally, AM mapping and spectrum reassignment by Meyer *et al* (1997) seems to provide some promise as a means of separating speakers under co-channel speech conditions. They suggest that the modulation maps are good models for human perceptual data, and by using a reassigned spectral approach, frequency resolution is increased.

References

1. Benincasa, D. S. and Savic, M. I., “Co-channel speaker separation using constrained nonlinear optimization,” Proc. IEEE ICASSP, pp:1195-1198, 1997.

2. Casajus Quiros, F. J. and Enriquez, P. F-C., “Real-time, loose-harmonic matching fundamental frequency estimation for musical signals,” Proc. IEEE ICASSP, Ii-221-II-224, 1994.

3. Chang, C., Ding, Z., Yau, S. F. and Chan, F. H. Y., "A matrix-pencil approach to blind separation of non-white sources in white noise," IEEE ICASSP, pp: IV-2485-IV-248, 1998.

4. Doval, B. and Rodet, X., “Estimation of fundamental frequency of musical sound signals,” Proc. IEEE ICASSP, pp:3657-3660, 1991.

5. Doval, B. and Rodet, X., “Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs,” Proc. IEEE ICASSP, pp:I-221-I-224, 1993.

6. Dorken, E. and Nawab, S. H., “Improved musical pitch tracking using principal decomposition analysis,” Proc. IEEE ICASSP, pp:II-217-II-210, 1994.

7. Kanjilal, P. P. and Palit, S., “Extraction of multiple periodic waveforms from noisy data,” Proc. IEEE ICASSP, pp:II-361-II-364, 1994.

8. Meyer, G. F., Plante, F., and Bethommier, “ Segregation of concurrent speech with the reassignment spectrum,” Proc. IEEE ICASSP, pp:1203-1206, 1997.

9. Morgan, D. P., George, E. B., Lee, L. T, and Kay, S. M., " Co-channel speaker separation," Proc. IEEE ICASSP, pp:828-831, 1995.

10. Openshaw, J. P. and Mason, J. S., “On the limitations of cepstral features in noise,”, Proc. IEEE ICASSP, pp: II-49-II-52, 1994.

11. Savic, M., Gao, H. and Sorensen, J. S., "Co-channel speaker separation based on maximum-likelihood deconvolution,", IEEE ICASSP, pp:I-25-I-28, 1994.

12. Weinstein, E., Feder, M., and Oppenheim, A. V., "Multi-Channel Signal Separation by Decorrelation," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. 1, No. 4, pp:405-413, Oct. 1993.

13. Wise, J. D., Caprio, J. R., and Parks, T. W., “Maximum likelihood pitch estimation,” IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-24, No. 5, pp:418-423, Oct. 1976.

14. Yen, K-C and Zhao, Y., “Co-channel speech separation for robust automatic speech recognition: stability and efficiency,” Proc. IEEE ICASSP, pp:859-862, 1997.

15. Yu, G., and Gish, H., “Identification of speakers engaged in dialog,” Proc. IEEE ICASSP, pp:II-383 – II-386, 1993.